Quick win!

Educate staff about bias in ML artificial intelligence

About: Staff that are working on ML systems should understand where and how bias can creep into data as well as algorithms. Promoting a culture of ethics and responsibility around artificial intelligence (Play #2 in the Mitigating Bias in AI playbook) requires this type of critical thinking and understanding.

Players: Project managers facilitate with their teams as participants.
Note: Responsible AI leads could also do this first with project managers.

What you will have after executing this 'quick win':

- A team that is primed for critical conversations about types of technical and non-technical biases
- A concrete example of your initiative to promote equity and responsibility technology development / to put your company's ethical and responsible AI principles into action
- A shared team-building experience that uncovers new perspectives among team members

Steps:

1. Schedule a brown bag lunch or mandatory meeting to facilitate a Case Study on bias in artificial intelligence with your team (see the following pages). We recommend scheduling 1.5 hours.
   a. Ask participating team members to review pages 21-36 of the Mitigating Bias in AI playbook prior.
   b. Include the link and directions in the invite to staff and let them know at least 1 week prior if possible.
2. Prepare your facilitation.
   a. Read pages 21-36 for more examples and context for your team.
   b. Review the case and facilitator notes.
3. Print out a copy of the case and the 'task' section for each participant.
4. Follow the tasks outlined in the Case Study and facilitate the discussion.
5. Debrief:
   a. Take notes for your personal record about the experience for you as a team leader
   b. Share with other project managers how it went.
   c. Remember this experience to reference in your own performance reviews and reporting.

COVID-19, Artificial Intelligence & Bias
Case Study

The Case Study below outlines a scenario related to bias in artificial intelligence (AI). As an organization, you are currently working to unlock the value of AI responsibly and equitably. As a group, we will:

- Think about how to make challenging decisions related to bias in artificial intelligence in real world scenarios
- Think about how our choices may be influenced by lived experiences

Description

Anita works at a healthcare technology company in San Francisco, MedCare Technology, Inc. She is a project manager working on AI systems with a background in engineering and an MBA degree. Under her leadership, her team created a machine learning system that predicts when patients will go into cardiac arrest. It extracts variables from health records of hospitalized patients at partner university hospitals. It is already used in several hospitals in the US. The system highlights when a patient becomes high risk to trigger an evaluation or to transfer the individual to an intensive care unit.

In March 2020, the novel coronavirus SARS-COV-2 (COVID-19) was declared a pandemic by the World Health Organization and shortly after a national emergency was declared in the United States regarding the outbreak. Given the scale of the pandemic, it was anticipated that hospitals in locations globally would be overrun and doctors overwhelmed, straining doctors' capacity to assess patient risk and make critical decisions timely and effectively.

Anita's company immediately kicked into gear wondering how it could adapt their cardiac arrest tool to help doctors and COVID-19 patients. They asked themselves, "How might we use AI to predict which patients will be high risk to COVID-19 complications? How might an early warning system help inform deployment and allocation of life-saving resources like ventilators?" The team was excited – many hospitals, particularly in New York, were already tearing at the seams with doctors attempting to support as many patients as possible and volunteers looking for direction. Her team could do something.

During the team's exploratory phase, Dr. Martin, a lung specialist doctor working at a large Bay Area hospital and advisor to the team, shared the following email:

Hi Team,

Anita asked me to share some information that might be relevant as you develop your AI model. Hope this helps and let me know if you have any follow up questions.

- *Patients that are higher risk tend to include: older patients and those with underlying medical conditions (e.g., obesity, type 2 diabetes, asthma). More from the Center for Disease Control and the underlying medical conditions is* here.
- *We have and can share data from chest CT scans of patients that received them at our hospital. CT scans use x-rays to identify COVID-19 signs and the extent of the virus in the lungs.*

- *We have and can share extensive health data from other COVID-19 patients we've had to date. Of course, all information shared will go through rigorous privacy and licensing procedures.*

Under Anita's guidance, your team gets to work.

Task:

1. Individually read the case study. *(Suggested time: 5 minutes)*
2. In groups of 2-3, answer the following questions: (Suggested time: 30 minutes)
    a. What concerns do you have about the data referenced by Dr. Martin? How might this data be biased? (Refer to the bias in AI map (pages 23-36 of the Mitigating Bias in AI playbook for examples and types of bias).

    b. Do you have any follow up questions for Dr. Martin? What other information or data would you like?

    c. What types of proxy variables would you like to include in the algorithmic model? Are there any ways that these proxies could embed bias? If so, how? (Hint: The medical system has a history of discrimination, an example of that is here)

    d. Are there other ways that bias could creep into the algorithm? (Refer to the bias in AI map for examples and types of bias).

    e. If the biases discussed so far were to lead to an inaccurate prediction, what impact(s) would that have?
3. Each group shares with the larger group their thoughts on the questions. (Suggested time: 20 minutes)
4. Questions for the larger group: (Suggested time: 15 minutes)
    a. What recommendations might you make to MedCare Technology Inc.'s CEO regarding how to proceed? What technical -- and non-technical -- approaches would you suggest? (Note: refer to pages 47-50 of the playbook for some ideas)

    b. What would you do if you discovered that the algorithm disproportionately mis-diagnosed one gender or ethnicity compared to others?
5. Share a slightly updated scenario by reading the following to participants: (Suggested time: 15 minutes)
    a. Let's imagine a different scenario… Instead of using the data to predict which patients are at high-risk for developing COVID-19 complications, Anita's team is instead tasked with determining what additional characteristics were common amongst those who were high-risk versus low-risk. When the team identified additional characteristics, they set up a meeting to work with Dr. Martin and her colleagues. They shared their findings with Dr. Martin and her colleagues to discuss if they've uncovered new patterns that may be actionable for doctors on the ground to help them better assess the risk of complications.

    b. How might this alternative address the concerns raised in Task 1 and 2a/2b above? What new concerns does this approach raise?
6. As a larger group, discuss the final question: (Suggested time: 5 minutes)
    a. How do you think your lived experiences informed your choices and decisions individually, and as a team?

Facilitator notes on how bias can be present in the data & algorithms

If participants need additional information or support to identify bias that can be present in the data or development of the algorithm, share some of the below information (which is also referenced in the Mitigating Bias in AI playbook).

Pathways to a biased dataset

Data points don't exist or data not disaggregated

- Immigrants are one of many communities of color hard-hit by COVID-19. Many immigrants are filling "essential" positions and are also at increased risk of complications or death from COVID-19 due to high rates of underlying chronic illnesses. However, many are not getting tested for fear of being deported. Data for trans and non-binary individuals may not exist. Lack of data on the most vulnerable isn't just true in the US – it's often even greater in poorer countries.
- The CDC released a race and sex breakdown of COVID-19 cases and deaths only in early April – pulling from parts of just 14 states. The data gaps matter because people react differently to viruses, vaccines and treatments, as illustrated in previous outbreaks (including SARS and Ebola). Official data that does exist around risk and mortality rates of COVID-19 is not sufficiently disaggregated by sex, race, or ethnicity, failing to tell an accurate story for millions of people. It was only in early April that an update released by the CDC contained a race and sex breakdown of data on COVID-19 cases and deaths – and it only pulled from hospital networks in parts of 14 states.
- Intersectional implications of the virus point to a need for disaggregating data by sex, as well as race/ethnicity, age, etc.

Datasets may exist but are biased

- COVID-19 infection rates are subject to a "vast undercount," by a factor of 50 or more. Also, patients are likely to get tested once they are already showing severe symptoms -- the data provided by Dr. Martin may not be representative of what early stage symptoms might look like for high-risk patients.
- Medical data is also being collected but only a certain subset of the population (often affluent, white) who can readily access the limited, expensive tests and medical procedures available (such as CT scans). Research shows that Black and Hispanic patients may be less likely to be recommended procedures such as CT scans.
- Since data is being provided by a single hospital, that can overrepresent certain characteristics linked to location. For instance, socioeconomic status of patients, local air quality and other potentially contributing factors, etc.
- The alarming rates of the virus among Black Americans is likely to be rooted in long standing economic and health care inequalities. Even this data is still incomplete and/or subject to inconsistent classification.

Data labels are subjective or discriminatory

- The poor quality of existing data related to race and ethnicity can be linked to ambiguous racial / ethnic categorization which further obscures disparities -- for instance, it is unclear whether Black Hispanics (e.g., Black Dominicans or Black Puerto Ricans) are being filed under "Black" or "Hispanic", and some states are reporting "Hispanic" as a race even though it is listed as an ethnicity in other government forms.

Pathways to a biased algorithm

Purpose of the algorithm

- In terms of optimizing for a certain objective – what are the fairness constraints it should operate under? This is not such an easy question to answer.
- In the alternative setting where Anita's team works as a collaborator with Dr. Martin and her colleagues to determine if there are actionable insights that the model has learned… This approach has the distinct advantage of letting AI be a tool to aid physicians in their work, as opposed to serving as a tool for displacement of a tough ethical questions regarding how resources should be allocated based on a risk score. Also, by passing the discovered features back to the medical team, the doctors have the ability to see if these insights are plausible given medical science, or if they are possibly confounders that the model inadvertently discovered.

Algorithm inputs (dataset use & proxies / variables)

- AI systems to detect and diagnose COVID-19 building from existing, available data are proliferating. Early AI systems such as a COVID-19 CT scanning algorithm utilized data from hospitals with a paper documenting a study not disclosing how patients were selected with severe selection bias (patients were in hospitals and had taken expensive test results). These patients are more likely affluent, white individuals.
- If AI systems assessing COVID-19 complications and risks utilize pre-existing conditions these can be linked to different identities. E.g., Black Americans have higher incidence of diabetes, which is tied to greater risk for COVID-19.

Algorithm evaluation

- Are social group categories and impacts on them incorporated in evaluation?

Other resources:

- Smith, G. & Rustagi, I. (2020). The problem with COVID-19 artificial intelligence solutions & how to fix them. Stanford Social Innovation Review.
- For an example of how a team like Anita's can help doctors uncover new patterns to assess risks of complications, see this talk by Rich Caruana of Microsoft Research: Friends don't let friends deploy black-box AI models.
- Information on eCART.

This Case Study was developed by the Center for Equity, Gender & Leadership (EGAL) at UC Berkeley Haas School of Business. It was written by Genevieve Smith with contributions from Nitin Kohli, Ishita Rustagi and Nandika Donthi. It is a resource of EGAL's playbook on Mitigating Bias in Artificial Intelligence. It is an accompanying resource of Mitigating Bias in Artificial Intelligence: An Equity Fluent Leadership Playbook.